Метод опорных векторов (SVM)

Метод опорных векторов (англ. *support vector machine*, *SVM*) — один из наиболее популярных методов обучения, который применяется для решения задач классификации и регрессии. Основная идея метода заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом. Алгоритм работает в предположении, что чем больше расстояние (зазор) между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше будет средняя ошибка классификатора.

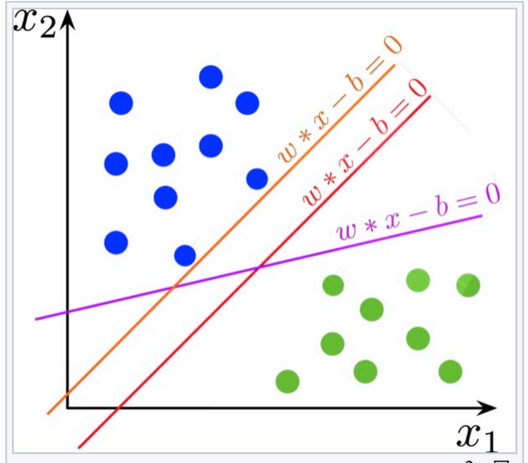
Метод опорных векторов в задаче классификации

Рассмотрим задачу бинарной классификации, в которой объектам из $X = \mathbb{R}^n$ соответствует один из двух классов $Y = \{-1, +1\}$.

Пусть задана обучающая выборка пар "объект-ответ": $T^\ell = (\vec{x}_i, y_i)_{i=1}^\ell$. Необходимо построить алгоритм классификации $a(\vec{x}) : X \to Y$.

Разделяющая гиперплоскость

В пространстве $\mathbb{R}^n$ уравнение $\langle \vec{w}, \vec{x} \rangle - b = 0$ при заданных $\vec{w}$ и b определяет гиперплоскость — множество векторов $\vec{x} = (x_1, \ldots, x_n)$, принадлежащих пространству меньшей размерности $\mathbb{R}^{n-1}$. Например, для $\mathbb{R}^1$ гиперплоскостью является точка, для $\mathbb{R}^2$ — прямая, для $\mathbb{R}^3$ — плоскость и т.д. Параметр $\vec{w}$ определяет вектор нормали к гиперплоскости, а через $\frac{b}{\|\vec{w}\|}$ выражается расстояние от гиперплоскости до начала координат.



Примеры разделяющих гиперплоскостей в $\mathbb{R}^2$

Гиперплоскость делит $\mathbb{R}^n$ на два полупространства: $\langle \vec{w}, \vec{x} \rangle - b > 0$ и $\langle \vec{w}, \vec{x} \rangle - b < 0$.

Говорят, что гиперплоскость разделяет два класса C_1 и C_2, если объекты этих классов лежат по разные стороны от гиперплоскости, то есть выполнено либо

$$\begin{cases} \langle \vec{w}, \vec{x} \rangle - b > 0, & \forall x \in C_1 \\ \langle \vec{w}, \vec{x} \rangle - b < 0, & \forall x \in C_2 \end{cases}$$

либо

$$\begin{cases} \langle \vec{w}, \vec{x} \rangle - b < 0, & \forall x \in C_1 \\ \langle \vec{w}, \vec{x} \rangle - b > 0, & \forall x \in C_2 \end{cases}$$

Линейно разделимая выборка

Пусть выборка линейно разделима, то есть существует некоторая гиперплоскость, разделяющая классы -1 и $+1$. Тогда в качестве алгоритма классификации можно использовать линейный пороговый классификатор:

$$a(\vec{x}) = sign(\langle \vec{w}, \vec{x} \rangle - b) = sign\left(\sum_{i=1}^{\ell} w_i x_i - b\right)$$

где $\vec{x} = (x_1, \ldots, x_n)$ — вектор значений признаков объекта, а $\vec{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ и $b \in \mathbb{R}$ — параметры гиперплоскости.

Но для двух линейно разделимых классов возможны различные варианты построения разделяющих гиперплоскостей. Метод опорных векторов выбирает ту гиперплоскость, которая максимизирует отступ между классами:

Отступ (англ. *margin*) — характеристика, оценивающая, насколько объект "погружён" в свой класс, насколько типичным представителем класса он является. Чем меньше значение отступа M_i, тем ближе объект $\vec{x}_i$ подходит к границе классов и тем выше становится вероятность ошибки. Отступ M_i отрицателен тогда и только тогда, когда алгоритм $a(x)$ допускает ошибку на объекте $\vec{x}_i$.

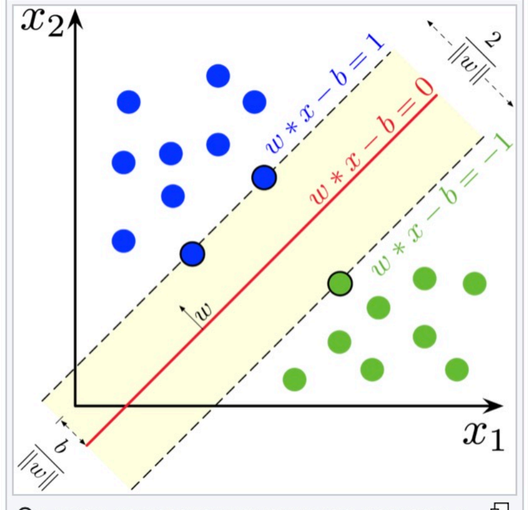
Для линейного классификатора отступ определяется уравнением: $M_i(\vec{w}, b) = y_i(\langle \vec{w}, \vec{x}_i \rangle - b)$

Если выборка линейно разделима, то существует такая гиперплоскость, отступ от которой до каждого объекта положителен:

$$\exists \vec{w}, b: \; M_i(\vec{w}, b) = y_i(\langle \vec{w}, \vec{x}_i \rangle - b) > 0, \; i = 1 \dots \ell$$

Мы хотим построить такую разделяющую гиперплоскость, чтобы объекты обучающей выборки находились на наибольшем расстоянии от неё.

Заметим, что при умножении $\vec{w}$ и b на константу $c \neq 0$ уравнение $\langle c\vec{w}, \vec{x} \rangle - cb = 0$ определяет ту же самую гиперплоскость, что и $\langle \vec{w}, \vec{x} \rangle - b = 0$. Для удобства проведём нормировку: выберем константу c таким образом, чтобы $\min M_i(\vec{w}, b) = 1$. При этом в каждом из двух классов найдётся хотя бы один "граничный" объект обучающей выборки, отступ которого равен этому минимуму: иначе можно было бы сместить гиперплоскость в сторону класса с большим отступом, тем самым увеличив минимальное расстояние от гиперплоскости до объектов обучающей выборки.



Оптимальная разделяющая гиперплоскость в $\mathbb{R}^2$

Обозначим любой "граничный" объект из класса $+1$ как $\vec{x}_+$, из класса -1 как $\vec{x}_-$. Их отступ равен единице, то есть

$$\begin{cases} M_+(\vec{w}, b) = (+1)(\langle \vec{w}, \vec{x}_+ \rangle - b) = 1 \\ M_-(\vec{w}, b) = (-1)(\langle \vec{w}, \vec{x}_- \rangle - b) = 1 \end{cases}$$

Нормировка позволяет ограничить разделяющую полосу между классами: $\{x : -1 < \langle \vec{w}, \vec{x}_i \rangle - b < 1\}$. Внутри неё не может лежать ни один объект обучающей выборки. Ширину разделяющей полосы можно выразить как проекцию вектора $\vec{x}_+ - \vec{x}_-$ на нормаль к гиперплоскости $\vec{w}$. Чтобы разделяющая гиперплоскость находилась на наибольшем расстоянии от точек выборки, ширина полосы должна быть максимальной:

$$\frac{\langle \vec{x}_+ - \vec{x}_-, \vec{w} \rangle}{\|w\|} = \frac{\langle \vec{x}_+, \vec{w} \rangle - \langle \vec{x}_-, \vec{w} \rangle - b + b}{\|w\|} = \frac{(+1)\Big(\langle \vec{x}_+, \vec{w} \rangle - b\Big) + (-1)\Big(\langle \vec{x}_-, \vec{w} \rangle - b\Big)}{\|w\|} =$$

$$= \frac{M_+(\vec{w}, b) + M_-(\vec{w}, b)}{\|w\|} = \frac{2}{\|w\|} \to \max \;\Rightarrow\; \|w\| \to \min$$

Это приводит нас к постановке задачи оптимизации в терминах квадратичного программирования:

$$\begin{cases} \|\vec{w}\|^2 \to \min\limits_{w,b} \\ M_i(\vec{w}, b) \geq 1, \quad i = 1, \dots, \ell \end{cases}$$

Линейно неразделимая выборка

На практике линейно разделимые выборки практически не встречаются: в данных возможны выбросы и нечёткие границы между классами. В таком случае поставленная выше задача не имеет решений, и необходимо ослабить ограничения, позволив некоторым объектам попадать на "территорию" другого класса. Для каждого объекта отнимем от отступа некоторую положительную величину ξ_i , но потребуем чтобы эти введённые поправки были минимальны. Это приведёт к следующей постановке задачи, называемой также *SVM с мягким отступом* (англ. *soft-margin SVM*):

$$\begin{cases} \frac{1}{2}\|\vec{w}\|^2 + C\sum\limits_{i=1}^{\ell}\xi_i \to \min\limits_{w,b,\xi} \\ M_i(\vec{w},b) \geq 1 - \xi_i, \quad i = 1,\dots,\ell \\ \xi_i \geq 0, \quad i = 1,\dots,\ell \end{cases}$$

Мы не знаем, какой из функционалов $\frac{1}{2}\|\vec{w}\|^2$ и $\sum\limits_{i=1}^{\ell}\xi_i$ важнее, поэтому вводим коэффициент C, который будем оптимизировать с помощью кросс-валидации. В итоге мы получили задачу, у которой всегда есть единственное решение.

Заметим, что мы можем упростить постановку задачи:

$$\begin{cases} \xi_i \geq 0 \\ \xi_i \geq 1 - M_i(\vec{w},b) \\ \sum\limits_{i=1}^{\ell}\xi_i \to \min \end{cases} \Rightarrow \begin{cases} \xi_i \geq \max(0, 1 - M_i(\vec{w},b)) \\ \sum\limits_{i=1}^{\ell}\xi_i \to \min \end{cases} \Rightarrow \xi_i = (1 - M_i(\vec{w},b))_+$$

Получим эквивалентную задачу безусловной минимизации:

$$\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{\ell}\big(1 - M_i(\vec{w},b)\big)_+ \to \min_{w,b}$$